Chloroflexi CL500-11 populations that predominate deep lake hypolimnion bacterioplankton rely on nitrogen-rich DOM metabolism and C1 compound oxidation.

Vincent J. Denef [1,#], Ryan S. Mueller [2], Edna Chiang [1],
James R. Liebig [3], Henry A. Vanderploeg [3]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109; [2] Department of Microbiology, Oregon State University, Corvallis, OR 97331; [3] NOAA Great Lakes Environmental Research Laboratory, Ann Arbor, MI 48108.

[#]Corresponding author: 1141 Kraus Natural Science, 830 N. University, Ann Arbor, MI 48109
vdenef@umich.edu, Phone: +1 (734) 764 6481, Fax: +1 (734) 763 0544

Author contributions: VJD designed the study, EC, JRL, and HAV performed experiments, VJD, RSM, HAV, and EC analyzed the data, and VJD wrote the paper.

Short title: Genomic traits of the CL500-11 lineage.

Submitted to AEM on September 14, 2015

Abstract.

The *Chloroflexi* CL500-11 clade predominates bacterial biomass in oxygenated hypolimnia of deep lakes worldwide, including the world's largest freshwater system, the Laurentian Great Lakes. Traits that allow CL500-11 to thrive and its biogeochemical role in these environments are currently unknown. Here, we found that a CL500-11 population was mostly present in off-shore waters along a transect in ultra-oligotrophic Lake Michigan (a Laurentian Great Lake). It occurred throughout the water column in spring, and only in the hypolimnion during summer stratification, contributing up to 18.1 % of all cells. Genome reconstruction from metagenomic data suggested an aerobic, motile, heterotrophic lifestyle with additional energy gained through carboxidovory and methylovory. Comparisons to other available streamlined freshwater genomes revealed that CL500-11 contains a disproportionate number of cell wall/capsule biosynthesis genes and the most diverse DOM substrate uptake spectrum, particularly for peptides. *In situ* expression patterns indicate the importance of DOM uptake and protein/peptide turnover, as well as Type I and Type II carbon monoxide dehydrogenase and flagellar motility. Location in the water column influenced expression patterns most, marked by increased bacteriorhodopsin expression and a response to oxidative stress in surface compared to deep waters. While carrying multiple adaptations to an oligotrophic lifestyle, investment in motility, large cell size, and its distribution in both oligotrophic and mesotrophic lakes indicate the ability to thrive under conditions where resources are more plentiful. Our data indicate that CL500-11 plays an important role in nitrogen-rich DOM mineralization in the extensive deep lake hypolimnion habitat.

47

48 **Introduction.**

49 Freshwater lakes are disproportionally active sites of carbon cycling relative to the
50 surface area they cover due to strong linkages to the surrounding land from which they receive
51 inorganic nutrients as well as organic carbon (1,2). Of the estimated 1.9 Pg of terrestrial organic
52 carbon freshwater systems process per year, nearly half is respired by bacteria (3-5). When
53 including soil dissolved organic carbon out-gassing, net freshwater carbon emissions are of the
54 same order of magnitude as net oceanic uptake (2). While photochemical mineralization of
55 organic carbon can predominate in lake habitats with high levels of photoactive radiation (6),
56 bacterial contributions to dissolved organic matter (DOM) mineralization are important as well
57 (7).

58 Yet, linkages between the metabolism of organic carbon and specific populations remains
59 limited, particularly in the less-studied hypolimnia of lakes, even for ubiquitous and highly
60 abundant taxa, due to challenges to isolate representatives of these taxa (8). In recent years, use
61 of culture-independent methods has provided insights into the metabolic potential of some key
62 heterotrophic freshwater lineages, such as LD12, the freshwater sister clade to marine SAR11
63 (*Alphaproteobacteria*; (9)), acI (*Actinobacteria*; (10,11)), and *Polynucleobacter*
64 (*Betaproteobacteria*; (12,13). A common feature of these ubiquitous and abundant heterotrophic
65 freshwater taxa is a streamlined genome ranging between 1 and 2.5 Mbp with a variety of genes
66 encoding the ability to uptake and metabolize DOM.

67 We currently have no information regarding the geochemical role of the CL500-11
68 lineage of the *Chloroflexi*. *Chloroflexi* are rarely observed in freshwater pelagic zones when
69 oxygen is plentiful (8), but CL500-11-like populations are a notable exception and are emerging

as a taxon restricted to lakes that are deep enough to maintain low temperatures (<10 °C) in the hypolimnion after stratification (14). This group was first observed in Crater Lake, OR, where it is abundant throughout the year, contributing up to 50% of all cells in the deep (15,16). Since its initial discovery, similar predominance has been observed in deep lakes around the world, including in Western Europe, East Asia, and most recently in two of the largest freshwater lakes in the world, Lake Superior and Lake Huron (14,17). It would be worthwhile to know more about the role of CL500-11 populations in deep lake hypolimnia as (1) a large proportion of all surface freshwater is contained in the hypolimnia of deep lakes, and (2) CL500-11 contribute a large proportion of all bacterial biomass in these habitats through a combination of numerical dominance and large cell size.

In this study, we analyzed the occurrence of a CL500-11 lineage population along a near-shore to off-shore transect in Lake Michigan based on sequencing of the V4 region of the 16S rRNA gene as well as CARD-FISH. To understand its success and role in the deep lake habitat, we reconstructed a near-complete genomic dataset from Lake Michigan metagenomic data and identified functions in the CL500-11 genome that were overrepresented relative to other ubiquitous freshwater heterotrophs with streamlined genomes. To gain further insights into the *in situ* functioning of this organism we generated metatranscriptomic data and identified highly expressed functions, as well as differential expression patterns based on season and location in the water column.

Materials and methods

Physical and geochemical analyses. A plankton survey system (PSS) was continuously lowered and raised at ~0.25 m s^{-1} in a sinusoidal path from 1-2 m beneath the lake surface to 2-4

m above the bottom as the R/V Laurentian moved at ~1.8 m s^{-1} while logging data every 0.5 second (18). The PSS contained sensors mounted on a V-fin to measure chlorophyll *a* (Wet Labs ECO Fluorometer, Sea-Bird Scientific), photosynthetically active radiation (PAR; a 4-pi sensor (Model QSP2300, Biospherical Instruments Inc.), dissolved oxygen (SBE43, Sea-Bird Scientific), turbidity (Wet Labs ECO NTU, Sea-Bird Scientific), and temperature. The fluorometer output (volts) was converted to derived chlorophyll *a* concentrations by regression between fluorometer output and laboratory chlorophyll *a* measurements (18). Replicate samples were analyzed for dissolved organic carbon (DOC), particulate organic carbon (POC) and nitrogen (PON), total and particulate phosphorus (TP, PP), and soluble reactive phosphorus (SRP) according to NOAA GLERL standard operating procedures (19). Total dissolved phosphorus values (TDP) were calculated by subtracting PP from TP.

Bacterial samples. Water samples originated from 5 m below the surface and 2-5 m above the lake floor at the near-shore (43° 11' 17" N, 86° 20' 38" W; April 23, July 15, and September 23, 2013) and off-shore station (43° 11' 59" N, 86° 34' 11" W; April 23, July 16, and September 24, 2013) along the NOAA Lake Michigan Muskegon transect (Fig. 1). Water was collected using a 30 L Niskin bottle, and pre-filtered through 210 and 20 µm nitex mesh into 10 L carboys. Carboys, funnels, and mesh were bleach-cleaned, MilliQ water rinsed until no bleach odor remained, and twice with sample water. Pre-filtered water was sequentially filtered onto 3.0 µm polycarbonate filters and 0.22 µm polyethersulfone filter membranes (142 mm, Millipore) using a Masterflex I/P peristaltic pump (Cole Parmer) between settings 11-13. Filters were folded with bacterial biomass facing inwards and submersed into RNAlater (Ambion). Samples were stored at -20 °C on board and transferred to a -80 °C freezer within 48 h of sampling.

115 Sample filtering was limited to 10 minutes and all samples were stored in RNAlater within 20
116 minutes of sampling.

117 **DNA/RNA extraction.** Duplicate nucleic acid extractions from the same 142 mm filter
118 membrane were performed for each of the field samples using a modified AllPrep
119 DNA/RNA/miRNA Universal kit protocol (Qiagen) (20). Part of the RNA fractions were
120 converted to cDNA using the ProtoScript II First Strand cDNA Synthesis Kit (New England
121 BioLabs).

122 **16S rRNA gene sequencing and analysis.** DNA and cDNA was submitted to the Joint
123 Genome Institute for 16S rRNA gene amplicon sequencing targeting the V4 region of the 16S
124 rRNA gene (515F/806R universal primers) (21). Pooled libraries were sequenced on an Illumina
125 MiSeq sequencer, using v2 chemistry 2x250 (500 cycles) paired-end reads. RTA v1.17.28 and
126 MCS v2.2.0 software were used to generate data. A random subset of 40,000 read pairs were
127 used for each of the field data samples to reduce computational time. We used mothur v.1.34.3 to
128 generate the operational taxonomic unit (OTU, 97 % sequence similarity) table, rarefy data at a
129 subsampling that allowed inclusion of all samples (n = 4,500; certain samples contained >85%
130 chloroplast sequences). We used the MiSeq standard operating protocol accessed on Dec 17,
131 2014 using SILVA release 119 for alignment and classification (22,23). The OTU table was
132 imported into Excel and OTU count data for the OTU classified as *Anaerolineaceae* was
133 extracted for generating Figure 2. All data is available on the Joint Genome Institute's genome
134 data portal (http://genome.jgi.doe.gov/; Project IDs 1041195 and 1041198).

135 **Metagenomic sequencing and analyses.** The Joint Genome Institute generated
136 metagenomic data from field samples (0.22-3 µm fraction only) collected at the off-shore station
137 in spring (5 m (surface; IMG GOLD Analysis ID Ga0007769) and 108 m (deep; Ga0007770)),

138 summer (108 m (Ga0007777) and 35 m (chlorophyll maximum; Ga0007778)), and fall (108 m
139 only (no chlorophyll maximum present); Ga0007786). Paired end 150 bp reads were generated
140 (HiSeq2000) from ~ 200 bp fragment size Nextera (summer and fall) or TruSeq (spring) libraries
141 (Illumina). Metagenomic data from individual samples were assembled by the JGI pipeline.
142 Shortly, BBDuk (filterk=27, trimk=27) was used to remove Illumina adapters, Illumina artifacts,
143 and phiX, and quality-trim both ends to Q12. Reads with quality scores averaging less than 8
144 over the read before trimming, or with length under 40 bp after trimming, were discarded.
145 Remaining reads were mapped to human HG19 with BBMap, discarding all hits over 93%
146 identity. Trimmed, screened, paired-end Illumina reads were assembled using SOAPdenovo
147 v1.05 (24)(default settings) using multiple k-mers (81, 85, 89, 93, 97, 101). The contigs resulting
148 from each k-mer were de-replicated using in-house Perl scripts. Contigs smaller than 1800 bp
149 were assembled using Newbler v2.8 (Life Technologies) to generate larger contigs (flags: -tr, -
150 rip, -mi 98, -ml 80). Newbler contigs and SOAPdenovo contigs larger than 1800 bp were
151 combined using minimus 2 (25) (flags: -D MINID=98 -D OVERLAP=80). Exact duplicate and
152 fully contained sequences were removed using dedupe (github.com/datamade/dedupe).
153 Full 16S rRNA gene reconstruction was performed with EMIRGE (default settings
154 except -l 150 –i 200 –s 50 –j 1.0) using SILVA v119 database clustered at 97% using
155 USEARCH (26) as reference. The setting –j 1.0 allows for the highest possible phylogenetic
156 resolution of the analysis, as only sequences that are 100% identical are combined during the
157 iterative read recruitment process. Sequences matching 18S rRNA genes were excluded.
158 To achieve improved population genome reconstructions, we used GAM-NGS (27) to
159 sequentially merge sample-specific assemblies from deep samples at the off-shore station in
160 spring, summer, and fall (Table S2). All resulting contigs > 3,000 bp were clustered using ESOM,

based on the tetranucleotide frequency of the contigs (28). To help extract contigs of interest (i.e., *Chloroflexi*), contigs were phylogenetically classified by searching open reading frames identified by prodigal (29) against NCBI NR using RAPsearch2 (30). We expanded the putative *Chloroflexi* ESOM bin with contigs > 1,000 bp that had a RAPsearch2 *Chloroflexi* classification and reduced it to the CL500-11-LM sequence bin using Maxbin (31) and the mmgenome R library (32) (Table S2). The latter relied on read recruitment with bowtie2 (default settings) (33) using 10 million paired end reads from off-shore spring surface and bottom, and summer bottom water samples. CheckM analysis was performed to assess completeness and purity of the final *Chloroflexi* CL500-11-LM sequence bin (34) and of the closest sequenced isolate, *Anaerolinea thermophila* UNI-1 (35). The CL500-11-LM bin was submitted to the DOE Joint Genome Institute Integrated Microbial Genomes for annotation. To remove redundancy due to incomplete assembly merging, a self-blastp analysis of the CL500-11-LM proteins was performed to search for replicate contigs by identifying blocks of proteins (2 or more) that were duplicated at >99% sequence identity. Redundant proteins were removed from the genome.

We used STAMP to identify over- and underrepresented protein functions in CL500-11-LM relative to a combined set of 6 freshwater heterotroph genomes (listed in Table S6), *A. thermophila*, and the deep-water summer off-shore station metagenomic dataset (36). Analyses were performed using the cluster of orthologous groups of proteins (COG) and Pfam profiles obtained from the JGI IMG data web portal. All datasets were rarefied to the lowest number of profiles (typically the number present in CL500-11-LM). Two-sample statistical analysis to identify functions that were present at significantly different relative abundances in CL500-11-LM relative to the reference genomes or metagenome was performed in STAMP using the

183 Fisher's exact test for Pfam profiles (due to low number of occurrence of many Pfams), and the
184 Chi-square test for COGs.

185 **Metatranscriptomic sequencing and analyses.** The Joint Genome Institute generated
186 metatranscriptomic data from RNA extracted from the offshore station samples (0.22-3 µm
187 fraction only) collected in spring (5 m daytime (Ga0007745) and nighttime (Ga0007744) and
188 108 m daytime (Ga0007746)), summer (108 m nighttime (Ga0007753) and 35 m daytime
189 (Ga0007754)), and fall (108 m nighttime (Ga0007762)). rRNA was was removed from 10 ng of
190 total RNA using Ribo-Zero™ rRNA Removal Kit (Bacteria) (Epicentre). Stranded cDNA
191 libraries were generated using the Illumina Truseq Stranded RNA LT kit. The rRNA depleted
192 RNA was fragmented and reversed transcribed using random hexamers and SSII (Invitrogen)
193 followed by second strand synthesis. The fragmented cDNA was treated with end-pair, A-tailing,
194 adapter ligation, and 15 cycles of PCR. Sequencing (2 x 150 bp) was performed on the Illumina
195 HiSeq2000 sequencer using a TruSeq SBS sequencing kits, v3.

196 Raw sequences were (1) quality-trimmed to Q10 and adapter-trimmed using BBDuk
197 (options: ktrim=r k=25 mink=12 tpe=t tbo=t qtrim=r trimq=10 maq=10 maxns=3 minlen=50),
198 (2) filtered for process artifacts using BBDuk (options: k=16), (3) mapped against a trimmed
199 version of the Silva database to remove remaining rRNA reads using BBMap (options: fast=t
200 minid=0.90 local=t), and (4) human reads were removed using BBMap. Remaining reads were
201 recruited to the curated CL500-11-LM coding genes using bowtie2 (options --local -p 10 -D 5 -R
202 1 -N 0 -L 25 -i S,1,2.00)(33). For samples with high abundance of CL500-11, 10 million paired
203 end read pairs were used (summer and fall deep night), while all reads were used for the
204 remaining datasets (summer chlorophyll maximum day, and spring surface day and night and
205 spring deep day). Mapped reads were scaled to the dataset with the lowest number of recruited

reads according to (37) and differential expression was detected using DEseq2, which identifies differentially expressed genes based on empirical Bayes shrinkage of both gene count variance and the fold-change estimation, two problematic issues of count-based data with a large dynamic range such as transcriptomics data (38). To account for Type I error, correction of the p-values using the Benjamini and Hochberg false-discovery rate correction available in the DeSeq2 package was performed. Overrepresentation of functions among highly expressed genes (5%, 10%, and 25% highest expressed genes relative to the proportion of genes encoding the same function in the whole CL500-11-LM genome), as well as among differentially expressed genes between surface and deep samples, was determined applying a Chi-square test for COG categories and a Fisher's exact test for Pfam categories (using the STAMP program). The p-values were corrected using the Benjamini and Hochberg false discovery rate correction.

CARD-FISH. We used the CARD-FISH probes (CL500-11-specific HRP-labeled probe CLGNS-584 and the helper probe CLGNS-567h to augment signal strength) and protocol developed by Okazaki (14) on 0.22-20 µm water subsamples from the water sampled for DNA and RNA extractions. Filters were examined with fluorescent microscopy by taking a photo and counting the number of DAPI-stained and probe-tagged cells within the field; a minimum of 300 DAPI-stained cells (5-10 fields of view) were counted per sample. Cell size and proportion of cell pairs was determined for ~ 50 cells per sample using the ZEN software measurement tools (Zeiss, Inc). Because CL500-11 cells have a curved rod shape, they were measured from the middle of the cell's width, following the cell shape length-wise. We identified the proportion of all cells that were part of dividing cells (cell pairs) and measured the length of each member of the cell pair separately. For all samples, cells were measured in either DAPI or GFP field depending on the clarity of the cell boundaries.

229

230 **Results.**

231 *Spatiotemporal distribution of CL500-11.*

232 All samples originated from the Muskegon transect on Lake Michigan, a long-term
233 observatory operated by NOAA GLERL (Fig. 1). The water column was isothermal in spring
234 and stratified in summer and fall (Fig. 1D-F), remained oxygenated throughout the season (Fig.
235 1G-I), and mid-day PAR was high in surface waters and low in waters below 60-80 m (Fig. 1M-
236 O). N, P and C levels were highest in the near-shore in spring, and lowest in the off-shore deep
237 water (Fig. 1C). Phytoplankton concentration as measured by chlorophyll *a* was low except for
238 near-shore in spring and in the off-shore chlorophyll maximum in summer (Fig. 1J-L).

239 Ribosomal rRNA sequencing data from cDNA and DNA libraries were combined in one
240 analysis, as DNA from summer deep off-shore water samples failed to amplify for 16S iTag
241 analyses. After rarefaction, 306 OTUs were classified as *Chloroflexi* in our data, containing
242 1.8 % of all rarefied sequencing reads (7th most abundant phylum; *Proteobacteria* are the most
243 abundant phylum accounting for ~ 40% of all rarefied reads). Only one of these OTUs reached
244 relative abundance > 3% in any of the individual sample datasets. This OTU was classified at the
245 family level as *Anaerolineaceae* and contained ~75% of all *Chloroflexi* sequences identified in
246 the entire dataset. This population was primarily found in the free-living fraction of the deep off-
247 shore waters (Fig. 2), and contributed up to 20.7 % of all sequences in a given sample (i.e., 0.22
248 – 3 µm, fall, off-shore, deep sample). To obtain more resolved taxonomic information for this
249 OTU, we used EMIRGE to reconstruct small subunit ribosomal RNA genes from the summer
250 off-shore station hypolimnion 0.22-3 µm fraction metagenomic data. We reconstructed six
251 unique *Chloroflexi* 16S rRNA genes (reconstructed from reference sequences of two

252 *Anaerolineaceae*, two SL56 marine group bacteria, one KD4-96, and one *Roseiflexus*),
253 containing a total of 10.1 % of all reads used to reconstruct bacterial sequences > 500 nt. One
254 sequence was generated using 7.3 % of all recruited reads and shared 99.6 % identity to clone
255 CL500-11 from Crater Lake, Oregon, USA (GenBank AF316759; Table S1) and to a partial
256 sequence retrieved from Lake Biwa, Japan (GenBank AB686531).
257 While predominantly present in the 0.22-3 µm fraction, CL500-11-like 16S rRNA gene
258 sequences were detected in the 3-20 µm fraction in fall, and had nearly equal relative abundances
259 in the offshore surface waters in spring (Fig. 2). To obtain clues regarding this apparent shift in
260 cell size, we performed CARD-FISH experiments on three off-shore samples (0.22-20 µm
261 fraction): surface water in spring, and bottom water in summer and fall. CARD-FISH data
262 confirmed the patterns observed by 16S V4 iTag sequencing (Fig 2). The absolute abundance of
263 CL500-11-like cells varied from 3.1×10^4 ($\pm 1.5 \times 10^4$; 95 % CI) cells/ml in spring surface
264 waters, to $2.2 \pm 0.8 \times 10^4$ and $4.6 \pm 1.3 \times 10^4$ cells/ml in summer and fall deep water samples,
265 respectively. Average cell size was larger in spring surface waters (0.87 ± 0.06 µm; 90 % C.I.) as
266 compared to summer and fall (0.76 ± 0.04 µm)(Fig. S1). In addition, the proportion of cells that
267 appeared to be connected cell pairs in spring was larger than that found in summer and fall,
268 though not significantly (71.4 % vs 50.1 % of all cells; $p > 0.10$; Fig. 2).
269
270 *CL500-11-LM genome reconstruction and analysis.*
271 In an effort to maximize the genome completeness, we merged three Lake Michigan
272 metagenomic dataset assemblies, and subsequently used tetranucleotide and differential coverage
273 binning methods (Fig. S2-S4) to isolate contigs comprising the CL500-11-LM population
274 genomic dataset. Initially, we selected 503 contigs with an average GC% of 60.9 that encoded

2,579 protein-coding genes (Table S2). CheckM estimated the genome to be 84.5 % complete, with 0.8 % contamination not attributable to strain heterogeneity (Table S3). After removal of redundancy resulting from the assembly merging process, we reduced the CL500-11-LM genome to 398 contigs encoding 2,153 protein-coding genes. CheckM indicated this significantly reduced within-genome redundancy (Table S3). The closest fully sequenced relative (*A. thermophila* UNI-1) was estimated to be 93.6 % complete (checkM). Corrected for the *A. thermophila* CheckM completeness estimate gap, we estimate the CL500-11-LM bin to be 90.2 % complete.

The curated CL500-11-LM sequence bin contained genes for aerobic oxidative phosphorylation (Table S4) and most genes required for glycolysis and the Krebs cycle (Table S4), including one of two key enzymes for the glyoxylate shunt (malate synthase, Ga0063436_10633). Missing were orthologs for the enzymes catalyzing phosphorylation of glucose – though multiple sugar kinases are present – and the conversion of malate to oxaloactetate. We also identified multiple ABC transporters for sugars and peptides (Table S5-S6). No evidence was found for carbon fixation capability, but we did identify genes involved in C1 compound oxidation: (1) Type I and Type II carbon monoxide dehydrogenases, based on both motif and gene neighborhood analyses, (Table S7), and (2) part of the pathway for tetrahydrofolate-dependent oxidation of methanol, glycine, methylamines, and potentially betaine (four copies of trimethylamine:corrinoid methyltransferases (39)) (Table S8). Missing genes were present in multiple copies in the metagenomic dataset, though were not part of the *Chloroflexi* bins. No genes encoding photosynthetic capabilities were found. However, a proteorhodopsin gene was encoded on a contig located on the margin of the *Chloroflexi* ESOM bin (Fig. S3), and read mapping across multiple datasets supported its inclusion in the CL500-11-

LM bin (Fig. S4). The retinal biosynthesis pathway was only partially present, with genes encoding beta-carotene 15,15'-monooxygenase (*bcmo*) and Lycopene cyclase (*crt*Y) missing from the CL500-11-LM bin based on analysis of the gene annotation and a Blastp search with the corresponding protein sequences from *Pelagibacter ubique* HTCC1062 (e-value < 1e-5).

Sixty-five % of CL500-11-LM coding genes did not have an ortholog in the most closely related sequenced *Chloroflexi* isolate, *A. thermophila* UNI-1 (ortholog defined as reciprocal best hit with blastp score ratio > 0.3, which corresponds to 30 % identity across 70% of the length of the protein (40)). Significantly fewer genes for regulatory processes and carbohydrate metabolism were encoded in the genome of CL500-11-LM compared to *A. thermophila*, while genes for amino acid and coenzyme transport and metabolism and cell motility were overrepresented in CL500-11-LM relative to the same functions in *A. thermophila* (Fig. 3, Table S9). The latter was the result of the unique presence in CL500-11-LM of the flagellar apparatus and two methyl-accepting chemotaxis proteins. In line with limited regulatory mechanisms in CL500-11-LM, only five RNA polymerase sigma factors were identified (sigma-28, sigma-54, and three sigma-70), compared to twelve sigma factors and two anti-sigma factors in *A. thermophila* UNI-1.

Genes involved in cell motility, amino acid transport and metabolism, and cell wall and membrane biosynthesis were overrepresented in CL500-11-LM compared to a set of six other abundant freshwater heterotrophic bacteria with relatively streamlined genomes (< 2.5 Mbp; two AcI lineage genomes, two LD12 lineage genomes, and two *Polynucleobacter* (PnecC) genomes, Table S6, S9) and the off-shore station deep water summer metagenome (Fig. 3). A more detailed comparison using Pfam profiles indicated a variety of peptide and carbohydrate transporters, peptidases, methyltransferases, and glycosyltransferases to be overrepresented in

CL500-11-LM relative to its closest sequenced relative and the community it dominates (Fig. 3, Table S9). DOM substrate uptake resembles AcI most closely (Table S6), though is more diverse and marked by much higher levels of redundancy of di- and oligopeptide transporters (31 proteins vs. five for AcI). The multiple copies of CO dehydrogenase genes were another unique feature of CL500-11-LM.

CL500-11-LM in situ *expression patterns.*

We recruited metatranscriptomic reads to the CL500-11-LM coding sequences for all datasets where iTag analysis indicated the presence of CL500-11-LM (Table S5, S10). Average expression levels at off-shore stations (spring surface and deep, summer chlorophyll maximum and deep, fall deep) indicated overrepresentation among the top 5% most highly expressed genes relative to their prevalence in the whole CL500-11-LM genome of amino acid transport and metabolism (14.5 % of all highly expressed genes relative to 8.3 % of all CL500-11-LM genes, q = 0.15), particularly di/oligopeptide transporters (5.2 % vs. 0.5 %, q = 0.001), energy production (14.5 % vs. 4.5 %, $q = 2.4x10^{-5}$) which includes both the Type I and Type II carbon monoxide dehydrogenases, and protein turnover (10.0 % vs. 2.0 %, $q = 2.2x10^{-6}$; Table S11). Relatively few housekeeping genes were among the most highly expressed genes, though when including the top 10% and top 25% most highly expressed genes, these functions were more highly represented, as well as genes involved in motility (Table S11). Transport functions for DOM were overrepresented in the CL500-11-LM transcriptome relative to the prevalence of these functions in he CL500-11-LM genome (11.4 % of all mRNA reads (Table S6)).

The highest number of differentially expressed genes was identified when comparing deep samples to surface water samples, regardless of season (75 genes; Table S12). Only 67

genes were differentially expressed when excluding the spring deep sample, 60 of which were also identified as differentially expressed when including the spring deep sample. Comparing spring to summer samples, regardless of depth (spring surface (day and night) and deep relative to summer and fall chlorophyll maximum and deep), resulted in only 7 differentially expressed genes. Finally, comparing spring surface and deep samples to summer and fall chlorophyll maximum and deep resulted in 26 differentially expressed genes, 19 of which were shared with the spring surface to summer/fall deep comparison. This indicated that location in the water column, which at the offshore station mostly differed in the level of photoactive radiation available (Fig. 1), was the most influential factor for CL500-11-LM gene expression. More genes were more highly expressed in the surface as compared to the deep, including defense mechanisms, Fe-S assembly proteins (*suf* operon), Fe uptake, components of the electron chain, glycolysis and TCA cycle, proteorhodopsin and proteins involved in carotenoid biosynthesis (Table S12). Genes more highly expressed in the deep included multiple amino acid metabolic functions including tryptophan biosynthesis, Type III glutamine synthetase, and part of a glycine cleavage complex, and part of the Type II carbon monoxide dehydrogenase operon.

Discussion.

While high *Chloroflexi* levels are rarely reported in oxygenated freshwater lakes (8), the CL500-11 lineage of the *Chloflexi* is emerging as a highly abundant taxon in the hypolimnion of deep stratified lakes around the world (14,15). Our study adds Lake Michigan to this list of lakes, which means CL500-11 has now been shown to reach relative abundances up to 20 % in the hypolimnia of all three upper Great Lakes (Lake Superior, Huron, Michigan; (17)). Its morphology, comparative genomic, and metatranscriptomic analyses reveal adaptations typical

of both oligotrophic and copiotrophic lifestyles, in line with their occurrence across a broad productivity gradient (14).

A recent summary of conditions under which CL500-11 lineage bacteria are found to be abundant indicates it can thrive under both oligotrophic and mesotrophic conditions, as long as oxygen does not get depleted in the hypolimnion and hypolimnion temperatures remain below 10 °C (Table 1 in (14)). Compared to Lake Biwa, where the same CARD-FISH probe was used, the Lake Michigan peak relative abundance is similar, while remaining well below the 50% reported in Crater Lake (15,16). Total bacterial and CL500-11 cell numbers are about one order of magnitude lower in Lake Michigan than in Lake Biwa, most likely reflecting the lower levels of DOC and nutrients in Lake Michigan, which is oligotrophic (Fig. 1) in contrast to mesotrophic Lake Biwa (14). CL500-11 remain present at high numbers at the end of the spring mixed period in Lake Michigan while they become undetected in Lake Biwa (14), potentially due to differences in the mixing regime of these two lakes. All current observations of CL500-11 are consistent with an oxygenated cold-water niche of deep, though not necessarily large, lakes. The only findings that argue against cold adaptation were identifications of CL500-11-like sequences near hydrothermal vents in Lake Yellowstone. However, one study indicated multiple lines of evidence for dilution of the vent water samples with surrounding cold (< 10 °C) water, including the presence of multiple non-thermophilic microorganisms and the presence of significant levels of O_2 in the sample while vent water was anoxic (41). The other study only observed the CL500-11-like sequences in the coldest vent sample (~16 °C) (42).

Our metabolic analysis as well as recent findings of aminosugar incorporation by a CL500-11-like population points to a heterotrophic lifestyle (43). In a previous study of Lake Biwa, it has been shown that semi-labile DOM produced by phytoplankton in the epilimnion is

390 subsequently biologically remineralized in the hypolimnion (44). Considering its abundance,
391 CL500-11 may play an important role in this process. Comparisons of the reconstructed CL500-
392 11-LM genome to other streamlined genomes of abundant freshwater heterotrophs identified
393 traits that support this function and may help explain its success in the deep lake hypolimnion
394 habitat. While some of the compared genomes are even more streamlined than CL500-11-LM
395 (e.g., LD12 (45) and AcI lineages (10,11)), many similarities and some contrasts can be found.
396 Similar to other freshwater heterotrophs, CL500-11 dedicates ~3% of its genome to import
397 diverse DOM substrates (46), but stands out as having the most diverse substrate spectrum,
398 particularly for di- and oligopeptides. In line with this observation, the overrepresentation of
399 metabolic functions to metabolize these peptides was observed. The low number of carboxylic
400 acid transporters relative to LD12 and *Polynucleobacter* could be a reflection of the preference
401 of CL500-11 for water depths with lower PAR, where there is a lower availability of these
402 compounds, which have been shown to be produced through the interaction between UV and
403 DOC (47).

404 In addition to the genomic overrepresentation of transporters, the disproportionate
405 number of transporters among highly expressed genes confirmed the importance of transporters
406 to CL500-11-LM, particularly of peptide transport. The prevalence of transporters among highly
407 expressed genes contrasts to a recent metatranscriptomic survey of a mesotrophic reservoir that
408 showed few transporters among the top 5% expressed genes (48). However, our results are in
409 line with community-level transcript measurements in coastal ocean systems, where up to 13 %
410 of all sequences assigned to COGs were involved in transport, and about half of those were
411 involved in DOM transport (49). In the most oligotrophic areas of the open ocean, proteomics
412 experiments have indicated two thirds of all protein originated from transport systems in SAR11

413 (50). Strikingly, the apparent importance of the CL500-11 lineage in organic nitrogen
414 transformation matches past findings that in the hypolimnion of Lake Biwa mineralization of
415 DOM preferentially converts N-rich compounds (51).

416 Carboxidovory and likely methylovory in CL500-11-LM allow for additional energy
417 generation through the oxidation of CO and methyl groups, respectively. Methylovory has been
418 demonstrated for *Candidatus Pelagibacter ubique* (52) and we identified homologous genes in
419 the SAR11 freshwater sister clade LD12. Carboxidovory in SAR11 has been suggested as one
420 mechanism in which it gains an edge in oligotrophic environments, such as the Sargasso Sea.
421 Sargasso Sea DOC concentrations are 2-3 x lower than deep off-shore Lake Michigan water,
422 while Lake Michigan soluble reactive phosphorus levels are lower (53,54). Lack of data
423 precludes comparison of inorganic nitrogen levels, though freshwater systems are typically more
424 P- than N-limited (55). Evidence of methylovory (i.e. carbon monoxide dehydrogenase genes) on
425 the other hand was not identified in any of the other streamlined freshwater heterotrophs or
426 SAR11. A carbon monoxide dehydrogenase is present in copiotrophic and often abundant
427 freshwater *Limnohabitans* species (56) and has been observed as a highly expressed gene in a
428 mesotrophic reservoir (48). The one-carbon oxidation pathways may be beneficial in both
429 oligotrophic and more productive systems where CL500-11 can be similarly abundant (14).

430 In addition to the large cell size, the overrepresentation of cell envelope biosynthesis
431 genes and energetically costly investment in motility are features that characterize a bacterial
432 copiotrophic lifestyle (57). However, the relatively small genome size, limited investment in
433 transcriptional regulation (3.5 % of protein coding genes) and signal transduction (1.5 %), and a
434 small proportion of secreted (3.2 %) or membrane-bound proteins (19.4 %) are consistent with
435 adaptation to an oligotrophic lifestyle for CL500-11 (57). Regulatory mechanisms are

436 particularly reduced, even in comparison to other streamlined freshwater genomes. Relative to *A.*
437 *thermophila* UNI-1 60 % and 40 % fewer genes were involved in signal transduction and
438 transcription, respectively, compared to a prediction of 20 % fewer genes in both categories
439 based on genome size difference alone (3,166 vs. 2,153 protein coding genes) (58). Overall, the
440 limited genome content and sequence conservation between CL500-11-LM and the filamentous
441 thermophile *A. thermophila* (35) are in line with large differences in morphology and ecology.
442 The cell envelope provides the interface that helps determine the outcome of interactions
443 with viruses and predatory grazers, primarily nanozooplankton (59). Escape from grazers due to
444 specific cell surface structures (S-layer) has been shown for the AcI lineage (60). Multiple
445 envelope structures have been inferred from genomic data for *Chloroflexi* lineages, including S-
446 layers (61). Although composition of the cell envelope can contribute to multiple traits, the
447 overrepresentation of cell envelope biosynthesis genes in CL500-11-LM relative to all other
448 genomes may underpin surface structures that allow escape from grazing predators, allowing it to
449 maintain high relative abundance levels. Independent of cell envelope composition, increased
450 grazing resistance may be conferred by the large cell size and curved morphology (62), as well
451 as by the large proportion of cell pairs. It is notable that in filter feeding experiments using water
452 and invasive quagga mussel samples from Lake Michigan, CL500-11-LM was enriched after 3.5
453 hours of filter feeding (V.J. Denef, unpublished results).
454 Occurrence in the surface waters in spring corresponded with increased cell size, which
455 could explain the increased occurrence in the 3-20 µm fraction in spring. The larger cells and
456 potentially increased abundance of cells pairs in the spring, which has previously been suggested
457 as indicative of cell division (14), could indicate more favorable conditions and a higher growth
458 rate in spring (63). However, the metatranscriptomic data did not indicate increases in core

459 functions such as transcription, DNA replication, and cell division, but instead indicated
460 exposure to stress in the surface waters. Particularly, multiple genes indicate the presence of
461 increased oxidative stress relative to the deep (*suf* operon (64), thioredoxin, *ahp*C, several
462 proteases involved in damaged protein turnover (65)), which is likely considering the increased
463 photoactive radiation at the surface (66). Oxidative stress response mechanisms are commonly
464 found in other streamlined freshwater heterotroph genomes as well (10,12), though different
465 protein families appear to be involved in CL500-11 and other freshwater genomes.

466 In light of the observed stress response, the higher expression of proteorhodopsin in the
467 surface water, which is expected based on higher PAR, suggests that CL500-11 uses
468 proteorhodopsin as a means to increase survival under suboptimal conditions rather than increase
469 cell growth (67). However, up-regulation of electron chain components, glycolysis, and TCA
470 enzyme-encoding genes is similar to the response seen during proteorhodopsin-induced growth
471 in a marine flavobacterium (68). Considering its marginal classification as part of the CL500-11-
472 LM bin, further verification of the presence of proteorhodopsin in the genome of CL500-11 and
473 its ecological role will be required.

474 CL500-11-like populations predominate the hypolimnion of at least three of the five
475 Laurentian Great Lakes, which contain ~20% of the world's surface freshwater. In addition, they
476 have been identified in multiple deep and often large lakes around the world, and can likely be
477 found in most deep lakes that maintain a cold and oxygenated hypolimnion. Their numerical
478 abundance, large cell size, and low within-clade sequence variation (14) mean that populations
479 with similar traits as CL500-11-LM likely contribute a significant proportion of the world's
480 freshwater bacterial biomass. Our analyses suggest they play an important role in the
481 transformation of biologically derived organic matter, particularly nitrogen-rich DOM.

Funding information.

VJD was supported by the Community Sequencing Program (U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, supported under Contract No. DE-AC02-05CH11231) and the University of Michigan. EC was supported by an American Society for Microbiology Undergraduate Research Fellowship and the Beckman Scholars Program. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Acknowledgments.

We are grateful to the crew of the R/V Laurentian, NOAA Great Lakes Environmental Research Laboratory science staff, and Marian L. Schmidt and Ann McCarthy (University of Michigan) for sampling and nucleic acid extractions. We thank the anonymous reviewers for their help in improving our manuscript. The authors declare they do not have any competing financial interests in relation to this work.

References.

1. **Cole JJ, Prairie YT, Caraco NF, McDowell WH, Tranvik LJ, Striegl RG, Duarte CM, Kortelainen P, Downing JA, Middelburg JJ.** 2007. Plumbing the global carbon cycle: integrating inland waters into the terrestrial carbon budget. Ecosystems **10:** 172-185.

2. **Tranvik LJ, Downing JA, Cotner JB, Loiselle SA, Striegl RG, Ballatore TJ, Dillon P, Finlay K, Fortino K, Knoll LB.** 2009. Lakes and reservoirs as regulators of carbon cycling and climate. Limnol Oceanogr **54:** 2298-2314.

505 3. **Cole JJ, Findlay S, Pace ML.** 1988. Bacterial production in fresh and saltwater ecosystems: a
506 cross-system overview. Mar Ecol Prog Ser **43:** 1-10.
507 4. **Cotner JB, Biddanda BA.** 2002. Small players, large role: microbial influence on
508 biogeochemical processes in pelagic aquatic ecosystems. Ecosystems **5:** 105-121.
509 5. **Leigh McCallister S, Del Giorgio PA.** 2008. Direct measurement of the d13C signature of
510 carbon respired by bacteria in lakes: Linkages to potential carbon sources, ecosystem baseline
511 metabolism, and CO2 fluxes. Limnol Oceanogr **53:** 1204-1216.
512 6. **Cory RM, Ward CP, Crump BC, Kling GW.** 2014. Carbon cycle. Sunlight controls water
513 column processing of carbon in arctic fresh waters. Science **345:** 925-928.
514 7. **Cole JJ, Carpenter SR, Kitchell JF, Pace ML.** 2002. Pathways of organic carbon utilization
515 in small lakes: Results from a whole-lake 13C addition and coupled model. Limnol Oceanogr
516 **47:** 1664-1675.
517 8. **Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S.** 2011. A guide to the natural
518 history of freshwater lake bacteria. Microbiol Mol Biol Rev **75:** 14-49.
519 9. **Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon K,**
520 **Bertilsson S, Stepanauskas R, Andersson SG.** 2013. Single-cell genomics reveal low
521 recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol **14:** R130.
522 10. **Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas**
523 **R, Grossart HP, Woyke T, Warnecke F.** 2013. Metabolic potential of a single cell belonging to
524 one of the most abundant lineages in freshwater bacterioplankton. ISME J **7:** 137-147.
525 11. **Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J,**
526 **Dwulit-Smith J, Chan LK, Martinez-Garcia M, Sczyrba A, Stepanauskas R, Grossart HP,**
527 **Woyke T, Warnecke F, Malmstrom R, Bertilsson S, McMahon KD.** 2014. Comparative

528 single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria
529 lineage. ISME J **8:** 2503-2516.
530 12. **Hahn MW, Scheuerl T, Jezberová J, Koll U, Jezbera J, Šimek K, Vannini C, Petroni G,**
531 **Wu QL.** 2012. The passive yet successful way of planktonic life: genomic and experimental
532 analysis of the ecology of a free-living polynucleobacter population. PLoS One **7:** e32772.
533 13. **Garcia SL, Buck M, McMahon KD, Grossart HP, Eiler A, Warnecke F.** 2015.
534 Auxotrophy and intrapopulation complementary in the 'interactome' of a cultivated freshwater
535 model community. Mol Ecol **24:** 4449-4459.
536 14. **Okazaki Y, Hodoki Y, Nakano S.** 2013. Seasonal dominance of CL500-11 bacterioplankton
537 (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. FEMS Microbiol
538 Ecol **83:** 82-92.
539 15. **Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ.** 2001. Unusual
540 bacterioplankton community structure in ultra-oligotrophic Crater Lake. Limnol Oceanogr **46:**
541 557-572.
542 16. **Urbach E, Vergin KL, Larson GL, Giovannoni SJ.** 2007. Bacterioplankton communities
543 of Crater Lake, OR: dynamic changes with euphotic zone food web structure and stable deep
544 water populations. Hydrobiologia **574:** 161-177.
545 17. **Rozmarynowycz MJ.** 2014. Spatio-Temporal Distribution Of Microbial Communities In
546 The Laurentian Great Lakes. Ph.D. thesis. Bowling Green State University, Bowling Green, MI,
547 USA.
548 18. **Vanderploeg HA, Ludsin SA, Ruberg SA, Höök TO, Pothoven SA, Brandt SB, Lang**
549 **GA, Liebig JR, Cavaletto JF.** 2009. Hypoxia affects spatial distributions and overlap of pelagic
550 fish, zooplankton, and phytoplankton in Lake Erie. J Exp Mar Biol Ecol **381:** S92-S107.

551 19. **Nalepa TF, Fahnenstiel GL, McCormick MJ, Johengen TH, Lang GA, Cavaletto JF,**
552 **Goudy G.** 1996. Physical and chemical variables of Saginaw Bay, Lake Huron in 1991-93.
553 NOAA Technical Memorandum ERL GLERL 91.
554 20. **McCarthy A, Chiang E, Schmidt ML, Denef VJ.** 2015. RNA preservation agents and
555 nucleic Acid extraction method bias perceived bacterial community composition. PLoS One **10:**
556 e0121659.
557 21. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM,**
558 **Betley J, Fraser L, Bauer M.** 2012. Ultra-high-throughput microbial community analysis on the
559 Illumina HiSeq and MiSeq platforms. ISME J **6:** 1621-1624.
560 22. **Schloss PD, Westcott SL.** 2011. Assessing and improving methods used in operational
561 taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ
562 Microbiol **77:** 3219-3226.
563 23. **Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.**
564 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-
565 based tools. Nucleic Acids Res **41:** D590-D596.
566 24. **Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S,**
567 **Yang H, Wang J, Wang J.** 2010. De novo assembly of human genomes with massively parallel
568 short read sequencing. Genome Res **20:** 265-272.
569 25. **Sommer DD, Delcher AL, Salzberg SL, Pop M.** 2007. Minimus: a fast, lightweight
570 genome assembler. BMC Bioinformatics **8:** 64.
571 26. **Edgar RC,** 2010. Search and clustering orders of magnitude faster than BLAST.
572 Bioinformatics **26:** 2460-2461.

573 27. **Vicedomini R, Vezzi F, Scalabrin S, Arvestad L, Policriti A.** 2013. GAM-NGS: genomic
574 assemblies merger for next generation sequencing. BMC Bioinformatics **14 Suppl 7:** S6.
575 28. **Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF.**
576 2009. Community-wide analysis of microbial genome sequence signatures. Genome Biol **10:**
577 R85.
578 29. **Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC.** 2012. Gene and translation initiation
579 site prediction in metagenomic sequences. Bioinformatics **28:** 2223-2230.
580 30. **Zhao Y, Tang H, Ye Y.** 2012. RAPSearch2: a fast and memory-efficient protein similarity
581 search tool for next-generation sequencing data. Bioinformatics **28:** 125-126.
582 31. **Wu YW, Tang YH, Tringe SG, Simmons BA, Singer SW.** 2014. MaxBin: an automated
583 binning method to recover individual genomes from metagenomes using an expectation-
584 maximization algorithm. Microbiome **2:** 26.
585 32. **Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH.** 2013.
586 Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of
587 multiple metagenomes. Nat Biotechnol **31:** 533-538.
588 33. **Langmead B, Salzberg SL.** 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods
589 **9:** 357-359.
590 34. **Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW.** 2015. CheckM:
591 assessing the quality of microbial genomes recovered from isolates, single cells, and
592 metagenomes. Genome Res **doi: 10.1101/gr.186072.114:** .
593 35. **Sekiguchi Y, Yamada T, Hanada S, Ohashi A, Harada H, Kamagata Y.** 2003.
594 Anaerolinea thermophila gen. nov., sp. nov. and Caldilinea aerophila gen. nov., sp. nov., novel

filamentous thermophiles that represent a previously uncultured lineage of the domain Bacteria at the subphylum level. Int J Syst Evol Microbiol **53:** 1843-1851.

36. **Parks DH, Tyson GW, Hugenholtz P, Beiko RG.** 2014. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics **30:** 3123-3124.

37. **McMurdie PJ, Holmes S.** 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol **10:** e1003531.

38. **Love MI, Huber W, Anders S.** 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15:** 550.

39. **Ticak T, Kountz DJ, Girosky KE, Krzycki JA, Ferguson DJ.** 2014. A nonpyrrolysine member of the widely distributed trimethylamine methyltransferase family is a glycine betaine methyltransferase. Proc Natl Acad Sci U S A **111:** E4668-E4676.

40. **Rasko DA, Myers GS, Ravel J.** 2005. Visualization of comparative genomic analyses by BLAST score ratio. BMC Bioinformatics **6:** 2.

41. **Clingenpeel S, Macur RE, Kan J, Inskeep WP, Lovalvo D, Varley J, Mathur E, Nealson K, Gorby Y, Jiang H, LaFracois T, McDermott TR.** 2011. Yellowstone Lake: high-energy geochemistry and rich bacterial diversity. Environ Microbiol **13:** 2172-2185.

42. **Yang T, Lyons S, Aguilar C, Cuhel R, Teske A.** 2011. Microbial communities and chemosynthesis in yellowstone lake sublacustrine hydrothermal vent waters. Front Microbiol **2:** 130.

43. **Tada Y, Grossart HP.** 2014. Community shifts of actively growing lake bacteria after N-acetyl-glucosamine addition: improving the BrdU-FACS method. ISME J **8:** 441-454.

616 44. **Maki K, Kim C, Yoshimizu C, Tayasu I, Miyajima T, Nagata T.** 2010. Autochthonous
617 origin of semi-labile dissolved organic carbon in a large monomictic lake (Lake Biwa): carbon
618 stable isotopic evidence. Limnology **11:** 143-153.
619 45. **Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon**
620 **K, Bertilsson S, Stepanauskas R, Andersson SG.** 2013. Single-cell genomics reveal low
621 recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol **14:** R130.
622 46. **Salcher MM, Posch T, Pernthaler J.** 2013. In situ substrate preferences of abundant
623 bacterioplankton populations in a prealpine freshwater lake. ISME J **7:** 896-907.
624 47. **Bertilsson S, Tranvik LJ.** 2000. Photochemical transformation of dissolved organic matter
625 in lakes. Limnol Oceanogr **45:** 753-762.
626 48. **Tsementzi D, Poretsky R, Rodriguez-R LM, Luo C, Konstantinidis KT.** 2014. Evaluation
627 of metatranscriptomic protocols and application to the study of freshwater microbial
628 communities. Environ Microbiol Rep **6:** 640-655.
629 49. **Poretsky RS, Sun S, Mou X, Moran MA.** 2010. Transporter genes expressed by coastal
630 bacterioplankton in response to dissolved organic carbon. Environ Microbiol **12:** 616-627.
631 50. **Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, Carlson**
632 **CA, Smith RD, Giovanonni SJ.** 2009. Transport functions dominate the SAR11 metaproteome
633 at low-nutrient extremes in the Sargasso Sea. ISME J **3:** 93-105.
634 51. **Kim C, Nishimura Y, Nagata T.** 2006. Role of dissolved organic matter in hypolimnetic
635 mineralization of carbon and nitrogen in a large, monomictic lake. Limnol Oceanogr **51:** 70-78.
636 52. **Sun J, Steindler L, Thrash JC, Halsey KH, Smith DP, Carter AE, Landry ZC,**
637 **Giovannoni SJ.** 2011. One carbon metabolism in SAR11 pelagic marine bacteria. PLoS One **6:**
638 e23973.

639 53. **Carlson CA, Ducklow HW, Michaels AF.** 1994. Annual flux of dissolved organic carbon
640 from the euphotic zone in the northwestern Sargasso Sea. Nature **371:** 405-408.
641 54. **Wu J, Sunda W, Boyle EA, Karl DM.** 2000. Phosphate depletion in the western North
642 Atlantic Ocean. Science **289:** 759-762.
643 55. **Hecky RE, Kilham P.** 1988. Nutrient limitation of phytoplankton in freshwater and marine
644 environments: A review of recent evidence on the effects of enrichment1. Limnol Oceanogr **33:**
645 796-822.
646 56. **Zeng Y, Kasalický V, Šimek K, Koblížeka M.** 2012. Genome sequences of two freshwater
647 betaproteobacterial isolates, Limnohabitans species strains Rim28 and Rim47, indicate their
648 capabilities as both photoautotrophs and ammonia oxidizers. J Bacteriol **194:** 6302-6303.
649 57. **Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting**
650 **L, Ertan H, Johnson J.** 2009. The genomic basis of trophic strategy in marine bacteria. Proc
651 Natl Acad Sci U S A **106:** 15527-15533.
652 58. **Konstantinidis KT, Tiedje JM.** 2004. Trends between gene content and genome size in
653 prokaryotic species with larger genomes. Proc Natl Acad Sci U S A **101:** 3160-3165.
654 59. **Jürgens K, Matz C.** 2002. Predation as a shaping force for the phenotypic and genotypic
655 composition of planktonic bacteria. Antonie van Leeuwenhoek **81:** 413-434.
656 60. **Tarao M, Jezbera J, Hahn MW.** 2009. Involvement of cell surface structures in size-
657 independent grazing resistance of freshwater Actinobacteria. Appl Environ Microbiol **75:** 4720-
658 4726.
659 61. **Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams**
660 **KH, Tringe SG, Banfield JF.** 2013. Community genomic analyses constrain the distribution of

661 metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling.
662 Microbiome **1:** 22.
663 62. **Grujčić V, Kasalický V, Šimek K.** 2015. Prey-Specific Growth Responses of Freshwater
664 Flagellate Communities Induced by Morphologically Distinct Bacteria from the Genus
665 Limnohabitans. Appl Environ Microbiol **81:** 4993-5002.
666 63. **Schaechter M, Maaløe O, Kjeldgaard NO.** 1958. Dependency on medium and temperature
667 of cell size and chemical composition during balanced growth of Salmonella typhimurium. J Gen
668 Microbiol **19:** 592-606.
669 64. **Nachin L, Loiseau L, Expert D, Barras F.** 2003. SufC: an unorthodox cytoplasmic
670 ABC/ATPase required for [Fe-S] biogenesis under oxidative stress. EMBO J **22:** 427-437.
671 65. **Cabiscol E, Tamarit J, Ros J.** 2010. Oxidative stress in bacteria and protein damage by
672 reactive oxygen species. Int Microbiol **3:** 3-8.
673 66. **Blough NV, Zepp RG.** 1995. Reactive oxygen species in natural waters. *In* Foote CS,
674 Valentine JS (ed), Active oxygen in chemistry. Chapman and Hall, New York, NY.
675 67. **DeLong EF, Béjà O.** 2010. The light-driven proton pump proteorhodopsin enhances
676 bacterial survival during tough times. PLoS Biol **8:** e1000359.
677 68. **Kimura H, Young CR, Martinez A, DeLong EF.** 2011. Light-induced transcriptional
678 responses associated with proteorhodopsin-enhanced growth in a marine flavobacterium. ISME J
679 **5:** 1641-1651.

680 **Figures.**

681 **Figure 1: Muskegon transect location and physicochemical data from sampling sites.** [A,B]
682 Southern Lake Michigan with depth contours and transect location. [C] Available geochemical
683 data (N, P, C) from the sample sites: S refers to 5 m below surface, M to 35 m below surface
684 (deep chlorophyll maximum), and D to 80 m below the surface. Error bars represent the standard
685 error between duplicate measurements. (D-O) Profiles determined by the plankton survey system
686 tow (dotted lines indicate travel path) at the time of sampling. Black and white circles indicate
687 microbiological field sampling sites. All data was collected during the nighttime transect, except
688 for PAR (photo-active radiation) data. The low surface PAR and wavy shape of the PAR profiles
689 across the spring transect result from overcast conditions plus rain showers on parts of the
690 transect.

691 **Fig. 2: Spatiotemporal distribution of CL500-11-LM *Chloroflexi* along the Lake Michigan**
692 **Muskegon transect.** Relative abundance of the CL500-11-like population as assessed by 16S
693 V4 sequencing and CARD-FISH across time, along the Muskegon transect (near-shore (15 (m))
694 and off-shore (110 (m)) station), at different depths (surface (S), bottom (D), and deep
695 chlorophyll maximum (M)), and in different size fractions (particle-associated (PA) and free-
696 living (FL). For CARD-FISH data, the fraction of all CL500-11 cells observed as part of a cell
697 pair is indicated as well. Errors bars indicate 95% C.I.

698 **Fig. 3: Differential representation of protein functions in the genome of CL500-11-LM**
699 **relative to genomes of other abundant freshwater heterotrophs with streamlined genomes**
700 **(SFH; see list in Table S6), summer Lake Michigan off-shore deep water metagenomic data,**
701 **and *A. thermophila* UNI-1.** Summary of STAMP analyses based on COG and Pfam profiles.
702 The effect size is the difference in relative abundance of a protein function in each genome or

703 metagenomic dataset. The q-values are the modified p-values after multiple testing correction
704 based on the Benjamini-Hochberg false discovery rate. The complete data set is included in
705 Table S9.
706